US-PAT-NO: **6593956**

DOCUMENT-IDENTIFIER: US 6593956 B1

Potts in view of Baker

TITLE: Locating an audio source

---------- KWIC ---------

Abstract Text - ABTX (1):

A system, such as a video conferencing system, is provided which includes an image pickup device, an audio pickup device, and an audio source locator. The image pickup device generates image signals representative of an image, while the audio pickup device generates audio signals representative of sound from an audio source, such as speaking person. The audio source locator processes the image signals and audio signals to determine a direction of the audio source relative to a **reference point**. The system can further determine a location of the audio source relative to the **reference point**. The **reference point** can be a camera. The system can use the direction or location information to frame a proper camera shot which would include the audio source.

US Patent No. - PN (1):

6593956

Brief Summary Text - BSTX (2):

This invention relates to systems, including video conferencing systems, which determine a direction of an audio source relative to a **reference point**.

Brief Summary Text - BSTX (5):

In one general aspect, the invention features a system which includes an image pickup device, an audio pickup device, and an audio source locator. The image pickup device generates image signals representative of an image, while the audio pickup device generates audio signals representative of sound from an audio source. The audio source locator processes the image signals and audio signals to determine a direction of the audio source relative to a **reference point**.

Brief Summary Text - BSTX (7):

In yet another general aspect, the invention features a video conferencing system including microphones, a camera, a positioning device, a processor, and a transmitter. The microphones generate audio signals representative of sound from an audio source and the camera generates video signals representative of a video image. The positioning device is capable of positioning the camera, for example, for tilting, panning, or zooming the camera. The processor processes the video signals and audio signals to determine a direction of a speaker relative to a **reference point** and supplies control signals to the positioning device for positioning the camera to include the speaker in the field of view of the camera, the control signals being generated based on the determined direction of the speaker. The transmitter transmits audio and video signals, which can be the same as the audio and video signals used for locating the audio source, for video-conferencing.

Brief Summary Text - BSTX (8):

In another general aspect, the invention features a system including microphones, a camera, a positioning device, a processor, and a transmitter. The microphones generate audio signals representative of sound from an audio
source and the camera generates video signals representative of a video image.
The positioning device is capable of positioning the camera, for example, for tilting, panning, or zooming the camera. The processor processes the audio signals to determine a direction of a speaker relative to a **reference point** and supplies control signals to the positioning device for positioning the camera to include the speaker in the field of view of the camera, the control signals being generated based on the determined direction of the speaker. The transmitter transmits audio and video signals, which can be the same as the audio and video signals used for locating the audio source, for video-conferencing.

Brief Summary Text - BSTX (12):

An image of a face of a person who may be speaking is detected in a frame of
video. The image of the face is detected by identifying a region which has flesh tone colors in the frames of video and may represent a moving face which
is determined, for example, by comparing the frame of video with a previous frame of video. It is then determined whether size of the region having flesh tone colors corresponds to a pre-selected size, the pre-selected size representing size of a pre-selected standard face. If the region having flesh tone colors corresponds to a flesh tone colored non-human object, the region is
determined not to correspond to an image of a face. The direction of the face relative to the **reference point** is also determined.

Brief Summary Text - BSTX (13):

The audio source locator includes an audio based locator for determining an
audio based direction of the audio source based on the audio signals and a

video based locator for determining a video based location of an image in one
of the frames of video. The image may be the image of the audio source which
may be an object or a face of a speaking person. The audio source locator then
determines the direction of the audio source relative to the **reference point** based on the audio based direction and the video based location.

Brief Summary Text - BSTX (15):

The audio source locator determines an offset of the video based location of
the image from a predetermined **reference point** in a frame of video and modifies
the audio based direction, based on the offset, to determine the direction of the audio source relative to the **reference point**. In this manner, the audio source locator can, for example, correct for errors in determining the direction of the audio source because of mechanical misalignments in components
of the system.

Brief Summary Text - BSTX (18):

Audio source locator correlates the audio based direction detected based on the audio signals to the stored video based location of the image in a frame of video and modifies the audio based direction, based on the results of the correlation, to modify audio based direction to determine the direction of the audio source relative to the **reference point**. To do so, for example, audio source locator modifies its processing to improve its accuracy.

Brief Summary Text - BSTX (22):

The audio source locator can also determine the distance from the **reference**
point to the audio source. The audio based locator determines a distance from

the **reference point** to the audio source based on the audio signals while the video based locator determines another distance from the **reference point** to the
audio source based on an image associated with audio source. Audio source locator then determines a finalized distance based on the audio based distance
and the video based distance.

Brief Summary Text - BSTX (25):

Determining the direction and/or location of an audio source relative to a **reference point** based on both audio and video provides for a system of checks
and balances improving the overall performance of the automatic camera pointing
system.

Detailed Description Text - DETX (3):

Briefly, during operation, video conferencing system receives sound waves from a human speaker and converts them to audio signals. Video conferencing
system also captures video images of the speaker. Video conferencing system 10
uses the audio signals and video images to determine a location of the speaker(s) relative to a **reference point,** for example, camera 14 or the center of rotation of camera positioning device 16. Based on that direction, video conferencing system 10 can then pan, tilt, or zoom in or out camera 14 to obtain a better image of the speaker(s).

Detailed Description Text - DETX (10):

Having described in general terms video conferencing system 10, the operation of audio source locator 28 of audio and video signal processor 20 will now be described in detail. An audio based locator (or audio based detector) 70 receives audio signals 22 and determines the location of a speaker

(i.e an audio source) relative to the microphone array. Audio based locator 70
then generates a series of camera positioning directives with respect to panning, tilting, and zooming camera 14. These directives can be partly based
on face detection and location analysis performed by a video based locator (or
video based detector module) 60. Audio based locator 70 then supplies a camera
control module 80 with these camera positioning directives. After camera control module 80 moves camera 14 according to these camera positioning directives, video based locator 60 analyzes the images in video frames 24 received as digital signals and stored as digital data in a memory storage unit (not shown). Video based locator 60 detects human faces in the images and determines their position relative to a **reference point** in the frame of video in which they are detected. Camera control module 80 then correlates a detected video face with the detected audio speaker and uses that correlation to correct or prevent camera framing errors.

Detailed Description Text - DETX (11):

FIG. 4 is a flow chart of the operation of audio source locator 28. Video based locator 60 includes processing modules 102-110, while audio based locator
70 includes processing modules 112-118. Each of these processing modules will
be described in detail below. Briefly, a video face location module 102 analyzes video signals 24 to detect faces in a single video frame. A video offset/error measurement module 104 measures the offset of the location of the
detected faces from some pre-determined and static **reference point** (for example, the center of the video image) and/or a dynamic **reference point** (for
example, the currently detected speaker). A face location tracking module 106
correlates the detected faces from the current video frame to the detected faces in the previous video frames and hence tracks the detected faces

through
a series of frames. This tracking allows for obtaining a proper position of a speaker in a video frame who may be moving, as will be described below. To
perform this tracking, face location tracking module 106 creates and maintains
a track file for each detected face.

Detailed Description Text - DETX (29):

Referring back to FIG. 4, after video face location module 102 finishes executing, video offset/error measurement module 104 determines the offset of
detected faces in the camera view from a video coordinate **reference point**. The
reference point can be a fixed **reference point** (for example, the center of the
camera image or a frame of video) or a dynamic **reference point** (for example,
location of a speaker detected by audio based locator 70). In either case, for each detected face, video offset/error measurement module 104 computes the offset by determining the X-axis and Y-axis difference between the center of the detected face and **reference point**. Where the **reference point** is a location
of a speaker detected by audio based locator 70, audio based locator 70 first converts the coordinates of the **reference point** from the audio coordinate system to the video coordinate system (step 112). Video offset/error measurement module 104 then uses these converted values to calculate the offset.

Detailed Description Text - DETX (58):

If the two conditions are met for a particular frequency component during a particular time frame, then it is assumed that an onset condition is met with respect to that frequency. A cross-spectrum for the audio signals acquired by the pair of microphones during the time frame is generated with respect to each

such frequency component, and a cross-spectrum for the noise at each such frequency is subtracted to identify the audio received signals representative of the sequence of signals from the audio source. The audio cross-spectrum is
accumulated during a predetermined length of time. If at the end of the predetermined time period, non-zero values for at least a specified number of frequencies have been accumulated, then the accumulated cross-spectrum values
are then used to compute (that is, are transformed to) cross-correlation values. The cross-correlation values in turn are used to determine the time delay between signals arriving at the pair of microphones from the common source. These time delays are then used to determine the direction and bearing
angle of the audio source with respect to the microphones which are used to determine a location of the audio source source from a predetermined **reference**
point such as the camera).

Detailed Description Text - DETX (68):

In some embodiments, one error for which camera control module 80 can correct is the error due to misalignment between camera 14 and microphone array
12. Generally, audio based locator 70 uses an array of microphones 12 to determine the location of a speaker relative to an audio **reference point**. The accuracy of this determination partly depends on the accuracy of the alignment
of camera 14 with array of microphones 12 through camera positioning device 16.
However, camera 14 and array of microphones 12 may be misaligned because of
mistakes during the manufacturing process or as a matter of regular use of the
system. Therefore, the camera pointing directives from audio based locator 70
can result in an image in which the speaker is offset from a desired position on the frame (for example, the center of the frame), as shown in FIG. 14.

Claims Text - CLTX (1):

1. A system comprising: an image pickup device generating image signals representative of an image; an audio pickup device generating audio signals representative of sound from an audio source; an audio based locator for processing the audio signals to determine a direction of the audio source relative to a **reference point**; a video based locator including a video face location module that processes the image signals for pixels having flesh tone colors to identify an object in the direction of the audio source; and a pointing control that steers the image pickup device to frame the object in the direction of the audio source.

Claims Text - CLTX (8):

8. The system of claim 6 wherein the audio source locator determines the direction of the face relative to the **reference point**.

Claims Text - CLTX (14):

14. The system of claim 13 wherein audio based locator determines an offset
of the video based location of the image from a pre-determined **reference point**
in said one of the frames of video and modifying the audio based direction, based on the offset to determine the direction.

Claims Text - CLTX (15):

15. The system of claim 13 further comprising a memory unit storing a previously determined offset of a video based location of an image in a previous one of the frames of video from a pre-determined **reference point,** wherein the audio source locator modifies the audio based direction, based on
the stored offset, to determine the direction.

Claims Text - CLTX (16):

16. The system of claim 6 wherein the speaking person moves relative to the **reference point,** the audio source locator detecting the movement of the speaker and, in response to the movement, causing an increase in the field of view of the image pickup device.

Claims Text - CLTX (28):

28. The system of claim 1 wherein the location is characterized by the direction of the audio source relative to the **reference point** and a distance, determined by the audio source locator, from the **reference point** to the audio source.

Claims Text - CLTX (29):

29. The system of claim 28 wherein the image signals represent frames of video images, the audio based locator determines an audio based distance from the **reference point** to the audio source based on the audio signals, the video based locator determines a video based distance from the **reference point** to the audio source based on an image of the audio source in one of the frames of video, and the audio source locator determines the distance based on the audio based distance and the video based distance.

Claims Text - CLTX (30):

30. A method comprising the steps of: generating, at an image pickup device, image signal representative of an image; generating audio signals representative of sound from an audio source; processing the audio signals to determine a direction of the audio source relative to a **reference point**; processing the image signals to identify an object in the direction of the

audio source; processing the image signals for pixels having flesh tone colors; and steering the image pickup device to frame the object in the direction of the audio source.

Claims Text - CLTX (34):
34. The method of claim 32 further comprising the step of determining the direction of the face relative to the **reference point**.

Claims Text - CLTX (40):
40. The method of claim 30 wherein processing the image signals and audio
signals further includes determining an offset of the video based location of the object from a pre-determined **reference point** in said one of the frames of
video, and modifying the audio based direction, based on the offset, to determine the direction.

Claims Text - CLTX (41):
41. The method of claim 30 wherein processing the image signals and audio
signals further includes modifying, based on a previously determined offset of
a video based location of an image in a previous one of the frames of video from a pre-determined **reference point,** the audio based direction to determine
the direction.

Claims Text - CLTX (42):
42. The method of claim 30 wherein the object is a speaking person that moves relative to the **reference point,** and wherein processing the image signals
and audio signals further includes: detecting the movement of the speaking person, and causing, in response to the movement, an increase in the field of

view of the image pickup device.

Claims Text - CLTX (53):

53. The method of claim 32 wherein the location is characterized by the direction of the audio source relative to the **reference point** and a distance, determined by the audio source locator, from the **reference point** to the audio
source.

Claims Text - CLTX (54):

54. The method of claim 53 wherein the image signals represent frames of video images, the method further comprising: determining a audio based distance
from the **reference point** to the audio source based on the audio signals, determining a video based distance from the **reference point** to the audio source
based on an image of the audio source in one of the frames of video; and determining the distance based on the audio based distance and the video based
distance.